Contextual simulated annealing Q-learning for pre-negotiation of agent-based bilateral negotiations

Tiago Pinto[1,2], Zita Vale[2]

[1] GECAD – Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
[2] Institute of Engineering, Polytechnic of Porto (ISEP/IPP), Portugal
{tcp, zav}@isep.ipp.pt

Abstract. Electricity markets are complex environments, which have been suffering continuous transformations due to the increase of renewable based generation and the introduction of new players in the system. In this context, players are forced to re-think their behavior and learn how to act in this dynamic environment in order to get as much benefit as possible from market negotiations. This paper introduces a new learning model to enable players identifying the expected prices of future bilateral agreements, as a way to improve the decision-making process in deciding the opponent players to approach for actual negotiations. The proposed model introduces a con-textual dimension in the well-known Q-Learning algorithm, and includes a simulated annealing process to accelerate the convergence process. The proposed model is integrated in a multi-agent decision support system for electricity market players negotiations, enabling the experimentation of results using real data from the Iberian electricity market.

Keywords: Bilateral Contracts, Context Awareness, Electricity Markets, Reinforcement Learning.

1 Introduction

The Electricity Markets (EM) restructuring placed several challenges to governments and to the companies that are involved in generation, transmission, and distribution of electrical energy [1]. Due to fuel fossil related concerns, the penetration of renewable energy sources has grown. The considerable increase of distributed generation of intermittent nature, makes EM more competitive, and consequently encourage a decrease in electricity prices [2]. However, some recurrent problems must be considered, e.g. the dispatch ability, limitations in the power system network, and the integration of small producers in EM, among others [3]. In order to overcome these

This work has received funding from the European Union's Horizon 2020 research and innovation programme under project DOMINOES (grant agreement No 771066) and from FEDER Funds through COMPETE program and from National Funds through FCT under the project UID/EEA/00760/2019.

problems, some global solutions are being adopted, deeply changing the traditional market models. Nowadays there are several market models, some with clearing mechanisms based on offers optimization, such as most US EM [4]; and other based on symmetric auctions, as in most European countries [5]. However, electricity trade worldwide is also supported by means of bilateral contracts negotiation.

With the increase of the complexity and unpredictability in EM, increases the need to understand markets' mechanism and how the interaction between the players affects markets outcomes. Simulation and decision support tools have been increasingly used, including several modeling tools based on multi-agent software. Some relevant examples are the simulators EMCAS AMES, GAPEX and MASCEM [6].

Current tools are directed to the study of market mechanisms and interactions among participants, but are not suitable for supporting the decisions of negotiating players in obtaining higher profits in energy transactions. The common behavior of market players in bilateral contracts negotiation is mainly based on the definition of prices and quantities in energy transactions with each competitor. Hence, relevant information, concerning competitors' previous negotiations, can be used to improve the decision process, considering the characteristics of the negotiation. It is essential to consider the concept of context awareness, since it influences the prices and volumes to be negotiated. A review of context analysis mechanism of EM players is presented in [7], which proposes a methodology to define and analyze different negotiation contexts in EM.

This paper presents a learning method to support decisions of players in the pre-negotiation of bilateral contracts, allowing to identify the ideal negotiators to trade with, enhancing the outcomes of the negotiation process. This method is based on the application of reinforcement learning algorithm (RLA), namely the Q-Learning algorithm, to learn the contract price forecasting method that is the closest to reality. This algorithm also determines the best method for each context. The forecast scenarios are determined using different methods to identify the expected price for each amount of energy. However, no method presents a better performance than all others in every situation, only in particular cases and contexts [8]. Thus, these contract prices forecasting are subject to some error degree. Because of that, the quality of definition of the best forecast method is essential for supporting the decision process. Besides the contextual dimension introduced in the learning process, a Simulated Annealing (SA) process [9] is also included to enable accelerating the convergence of the learning process, especially when the number of observations is low.

2 Proposed Methodology

The proposed method uses a learning process based on the assessment of likelihood of occurrence of each alternative scenario of negotiation. Thus, this approach allows the supported player to be prepared for the negotiation scenario that is the most likely to occur and perform the action that generates the best results. Besides, the contextualization of the learning process is enabled, obtaining the expected negotiation scenarios that most reflect the current context.

2.1. Contextual Q-Learning

The bilateral contract price estimation approach is based on the application of the Q-Learning reinforcement learning algorithm, where an agent learns through trial and error. An agent operates in an environment conceptualized by a set of possible states, in which the agent can choose actions from a set of possible actions. Each time that the player performs an action, a reinforcement value is received, indicating the immediate value of the resulting state transition. Thus, the only learning source is the agents' own experience, whose goal is to acquire an actions policy that maximizes its overall performance [10].

The proposed methodology proposes an adaptation of the Q-Learning algorithm [11] to undertake the learning process. Q-Learning is a very popular reinforcement learning method. It is an algorithm that allows the autonomous establishment of an interactive action policy. It is demonstrated that the Q-Learning algorithm converges to the optimal proceeding when the learning *Q* state-action pairs is represented in a table containing the full information of each pair value [12]. The basic concept behind the proposed Q-Learning adaption is that the learning algorithm can learn a function of optimal evaluation over the whole space of context-scenario pairs (*c x s)*. This evaluation defines the *Q* confidence value that each scenario can represent the actual encountered negotiation scenario *s* in context *c*. The *Q* function performs the mapping as in (1):

$$Q\colon c\;x\;s\;\rightarrow\;U \tag{1}$$

where *U* is the expected utility value when selecting a scenario *s* in context *c*. The expected future reward, when choosing the scenario *s* in context *c,* is learned through trial and error according to (2):

$$Q_{t+1}(c_{t,}s_t) = Q_t(c_{t,}s_t) + \alpha(c_{t,}s_t)[r_{s,c,t} + \gamma \cdot U_t(c_{t+1}) - Q_t(c_{t,}s_t)] \tag{2}$$

where c_t is the kind of context when performing under scenario s_t at time *t:*

- $Q_t(c_{t,}s_t)$ represents the value of the previous iteration (each iteration represents each new contract established in the given scenario and context). Generally, the *Q* value is initialized to 0.
- $\alpha(c_{t,}s_t)$ $(0 < \alpha \leq 1)$ is the learning rate which determines the extent to which the newly acquired information will replace the old information (e.g. assuming a value of 0 learns nothing; on the other hand, a value of 1 represents a fully deterministic environment).
- $r_{s,c,t}$ is the reward, which represent the quality of the pair context-scenario *(c x s)*. It appreciates the positive actions with high values and negative with low values, all of them are normalized on a scale from 0 to 1. The reward *r* is defined in (3):

$$r_{s,c,t} = 1 - |RP_{c,t,a,p} - EP_{s,c,t,a,p}| \tag{3}$$

where $RP_{c,t,a,p}$ represents the real price that has been established in a contract with an opponent p, in context c, in time t, referring to an amount of power a; and $EP_{s,c,t,a,p}$ is the estimation price of scenario that corresponds to the same player, amount of power and context in time t. All r values are normalized in a scale from 0 to 1.

- γ $(0 \leq \gamma \leq 1)$ is the discount factor which determines the importance of future rewards. A value of 0 only evaluates current rewards, and higher values than 0 takes into account future rewards.
- $U_t(c_{t+1})$ is the estimation of the optimal future value which determines the utility of scenario s, resultant in context c. U_t is calculated as in (4):

$$U_t(c_{t+1}) = \max_{s} Q(c_{t+1}, s) \tag{4}$$

The Q-Learning algorithm is executed as follows:

- For each c and s, initialize $Q(c,s) = 0$;
- Observe new event (new established contract);
- Repeat until the stopping criterion is satisfied:
    - Select new scenario for current context;
    - Receive immediate reward $r_{s,c,t}$;
    - Update $Q(c,s)$ according to (2);
    - Observe new context c';
    - $c \rightarrow c'$.

After each update, all Q values are normalized according to the equation (5), to facilitate the interpretation of values of each scenario in a range from 0 to 1.

$$Q'(c,s) = \frac{Q(c,s)}{max[Q(c,s)]} \tag{5}$$

The proposed learning model assumes the confidence of Q values as the probability of a scenario in a given context. $Q(c,s)$ learns by treating a forecast error, updating each time a new observation (new established contract) is available again. Once all pairs context-scenario have been visited, the scenario that presents the highest Q value, in the last update, is chosen by the learning algorithm, to identify the most likely scenario to occur in actual negotiation.

2.2. Simulated annealing process

SA is an optimization method that imitates the annealing process used in metallurgy. The final properties of this substance depend strongly on the cooling schedule applied, i.e. if it cools down quickly the resulting substance will be easily broken due to an imperfect structure, if it cools down slowly the resulting structure will be well organized and strong. When solving an optimization problem using SA the structure of the substance represents a codified solution of the problem, and the temperature is used to determine how and when new solutions are perturbed and accepted. The algorithm is basically a three steps process: perturb the solution, evaluate the quality of the solution, and accept the solution if it is better than the previous one [13].

The two main factors of SA are the decrease of the temperature and the probability of acceptance. The temperature only decreases when the acceptance value is greater than a stipulated maximum. This acceptance number is only incremented when the probability of acceptance is higher than a random number, which allows some solutions to be accepted even if their quality is lower than the previous. When the condition of acceptance is not satisfied, the solution is compared to the previous one, and if it is better, the best solution is updated. At high temperatures, the simulated annealing method searches for the global optimum in a wide region; on the contrary, when the temperature decreases the method reduces the search area. This is done to try to refine the solution found in high temperatures. This is a good quality that makes the simulated annealing a good approach for problems with multiple local optima. SA, thereby, does not easily converge to solutions near the global optimum; instead this algorithm seeks a wide area always trying to optimize the solution. Thus, it is important to note that the temperature should decrease slowly to enable exploring a large part of the search space. The considered stopping criteria are: the current temperature and the maximum number of iterations. In each iteration is necessary to seek a new solution, this solution is calculated according to (6).

$$new\ solution = solution + S \times N(0,1) \tag{6}$$

solution in (1) refers to the previous solution, because this may not be the best found so far. $N(0,1)$ is a random number with a normal distribution, the variable S is obtained through (7).

$$S = 0.01 \times (upbound - lwbound) \tag{7}$$

$upbound$ and $lwbound$ are the limits of each variable, which prevent from getting out of the limits of the search problem.

The decisive parameters in SA's research are the decrease of temperature and the likelihood of acceptance. 4 variations of the SA algorithm have been implemented, combining different approaches for calculating these two components. It is expected that this will bring different results for different groups, as these components introduce a strong randomness in SA, which makes them reflect in the final results.

Table 1: Temperature and probability of acceptance calculation methods

group	temperature decreasing	probability of acceptance	ref.
1	$T_i = T_{i-1} \times \alpha$	$P = (2\pi T)^{-\frac{D}{2}} e^{\left(\frac{-\Delta x}{K \times T}\right)}$	[14]
2	$T_i = \frac{T_0}{i}$	$P = \frac{T_0}{(\Delta x^2 + T^2)^{\frac{(D+1)}{2}}}$	[14]
3	$T_i = T_0 e^{-ci^{\frac{1}{D}}}$	$P = \prod_{d=1}^{D} \frac{1}{2(\lvert y_d \rvert + Ti) \ln\left(1 + \frac{1}{T_i}\right)}$	[14]
4	$T_i = T_0 \times \alpha^i$	$T_i = \frac{1}{1 + e^{\frac{\Delta x}{T_{max}}}}$	[15]

where:

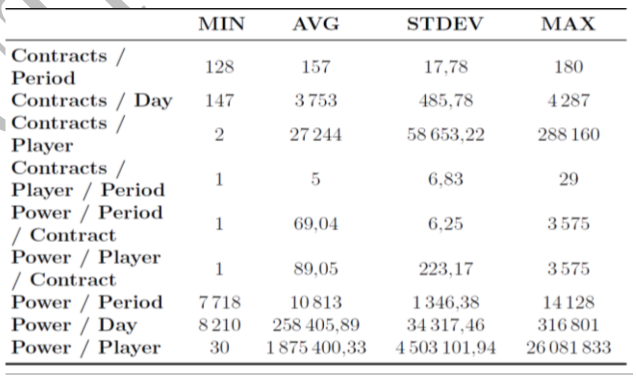
- $\alpha = 0.95$;
- i is the current iteration;
- $\Delta x = y(x^{max} - x^{i})$ is the difference between best solution and current solution;
- $K = 1$ is the Boltzmann constant ;
- $T_0 = 1$ is the initial temperature;
- D is the number of variables;
- $c = 0.1$;
- $|y_d|$ is the abs of solution current;
- $T_{min} = 1 \times 10^{-10}$;
- $acceptance_{max} = 15$.

3 Results and discussion

3.1. Case study characterization

This section presents a case study to demonstrate the performance of the proposed methodology. A historical database, concerning the past log of established contracts of different EM players, is used to apply the proposed methodology and assess its performance. The used data is based on real data extracted from MIBEL - the Iberian Electricity Market. The dataset can be consulted [16] and is composed by the executed physical bilateral contracts declared in the Spanish System Operator, in the period between 1 July 2007 and 31 October 2008 (16 months / 488 days). Each negotiation day is composed by 24 negotiation periods, in a total of 11712 periods. The negotiations were performed by 132 different players (88 Buyers and 44 Sellers) which established 1,797,996 contracts. Table 2 presents a detailed overview of the dataset.

Table 2: Dataset overview



	MIN	AVG	STDEV	MAX
Contracts / Period	128	157	17,78	180
Contracts / Day	147	3 753	485,78	4 287
Contracts / Player	2	27 244	58 653,22	288 160
Contracts / Player / Period	1	5	6,83	29
Power / Period / Contract	1	69,04	6,25	3 575
Power / Player / Contract	1	89,05	223,17	3 575
Power / Period	7 718	10 813	1 346,38	14 128
Power / Day	8 210	258 405,89	34 317,46	316 801
Power / Player	30	1 875 400,33	4 503 101,94	26 081 833

The distinct scenarios, which are the actions that the model may choose, refer to 5 contract price forecast methods, where there is an expected price for each amount of energy (from 1 until 10 MWh). The expected prices for the power amounts are calculated by several forecasting algorithms detailed in [17]. The context awareness is tested through 4 different contexts. The context analysis is carried out by a context analysis mechanism [7], which separates the historic data into different groups or contexts. 47% of the established contracts refer to Context 1, 8% refer to Context 2, 18% to Context 3 and 27% to Context 4.

The overall goal is to update the *Q* value of each forecast method (scenario) and context whenever there are new contracts. It is also important to test different combinations of input parameters, such as discount factor, learning rate and initial temperature; to analyze the evolution of *Q* values; and to have a suitable learning mechanism, which chooses the most likely forecast method to occur (i.e. the scenario with a lower forecast error in the current context).

Table 3 shows the comparison of the average error between the predicted price by each scenario and the actual verified price for each of the 4 considered scenarios. The error evaluation is measured using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Standard Deviation (STD). Using these prediction errors as basis it is possible to assess the quality of each scenario vs context pair, thus enabling to evaluate the quality of the learning process.

Table 3: Prediction error of the five considered scenarios in each of the four considered contexts

Context	Scenario	MAE	MAPE (%)	STD
1	1	13.94	18.43	16.34
	2	8.75	12.48	6.36
	3	3.26	4.12	2.36
	4	19.74	27.39	17.47
	5	12.89	17.62	9.20
2	1	3.67	5.26	4.62
	2	19.84	28.53	15.62
	3	3.82	5.48	4.89
	4	26.84	39.98	18.74
	5	10.31	14.53	8.38
3	1	3.91	7.52	4.93
	2	9.93	14.45	8.22
	3	3.74	7.16	4.53
	4	14.42	19.36	12.60
	5	6.22	9.36	6.83
4	1	5.46	8.63	7.31
	2	20.31	33.16	16.82
	3	24.17	38.28	18.45
	4	8.02	12.21	8.24
	5	15.16	21.75	13.82

From Table 3 it is visible that scenarios with lowest prediction error are: Scenario 1 for Contexts 2 and 4, and Scenario 3 for Contexts 1 and 3. These are the best scenarios (actions) to which the learning model should converge.

3.2. Results

Fig. 1 presents the heat maps showing the results quality (overall prediction errors) achieved by the proposed method when applied to each of the contexts independently. The heat maps include the combinations between the values of α and T_0. The dark green zones represent the combinations of α and T_0that present the best performance in each test, and the red zones represent the worst combinations.

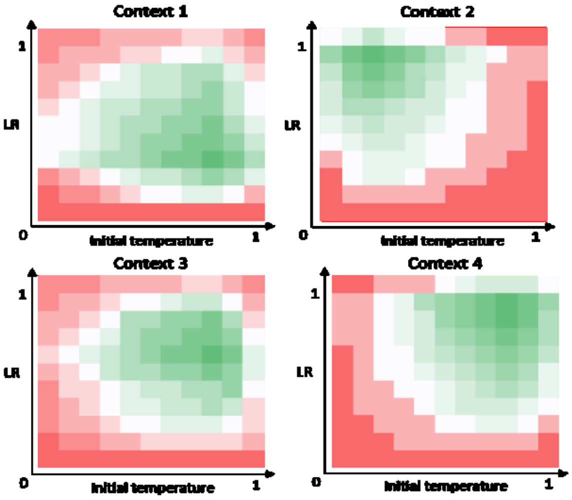


Figure 1: Sensitivity analysis results for the combination between α and T_0 in the different performed tests.

From Fig. 1 it is seen that the best combination of parameters depends greatly on the number of observations that refer to each of the contexts. In Context 1, being that with the greatest number of observations, the best results arrive when a large T_0 is set, together with a low α. This enables the learning process to learn more slowly, providing enough room for exploration before starting to exploit the best action. On the contrary, in Context 2, which is the context with the smaller number of observations a large α and small T_0 is required, so that the learning process converges more quickly. In Contexts 3 and 4, having a moderate number of observations, the tendency goes to an intermediate level of α associated to a rather large T_0, enabling a moderate learning process, with enough exploration before the final convergence. For illustrative purposes, Fig.2 shows the convergence process of the proposed model for Context 2, with and without SA.

Fig. 2 shows that, using the SA, the convergence to scenario 1 – the best action for context 2, is faster. There is less exploration, but the exploitation begins much sooner, which is important in contexts such as this one, in which the total number of observations is low.

Using the identified best parametrization, the results of the proposed methodology are compared to several benchmark reinforcement learning algorithms under the same simulation settings, namely the standard Q-Learning, Roth-Erev [18], UCB1 [19] and EXP3 [20]. Table 4 shows the global results, i.e. normalized confidence values (or Q values) in each of the 5 considered scenarios, in each of the 4 considered contexts. Table 5 shows the comparison of the average prediction errors resulting from the scenarios chosen in each iteration by the different algorithms in each context. This enables assessing the overall quality of the learning methods in each context. Note that it is not expected that the achieved error values match those achieved by the best scenarios themselves, as presented in Table 3, because due to the required exploration phase of the reinforcement learning algorithms several different scenarios, even if bad, must be tried, which results in an overall trial and error procedure. However, these average errors enable assessing the algorithms quality in terms of exploration vs exploitation balance, and their capability of converging to the best scenario, as shown by the confidence values in each scenario, as shown by Table 4.

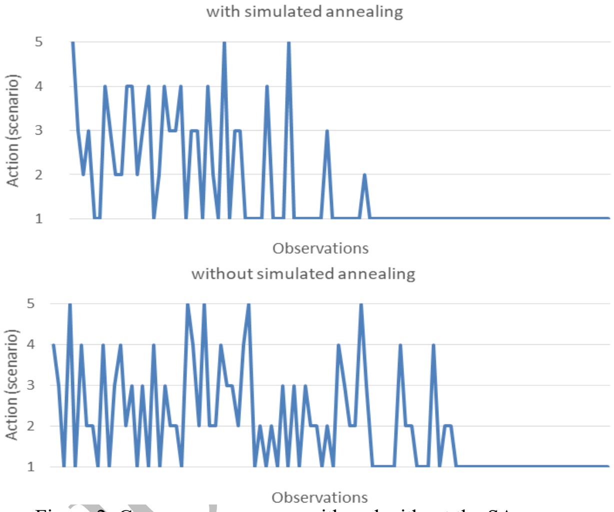
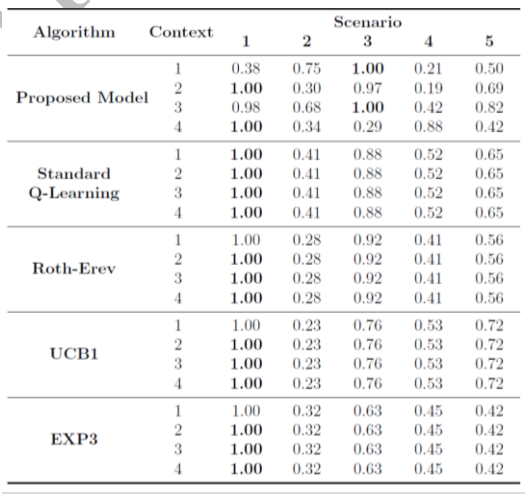


Figure 2: Convergence process with and without the SA process

Table 4: Comparative results between the proposed model and benchmark reinforcement learning algorithms



Algorithm	Context	Scenario 1	2	3	4	5
Proposed Model	1	0.38	0.75	**1.00**	0.21	0.50
	2	**1.00**	0.30	0.97	0.19	0.69
	3	0.98	0.68	**1.00**	0.42	0.82
	4	**1.00**	0.34	0.29	0.88	0.42
Standard Q-Learning	1	**1.00**	0.41	0.88	0.52	0.65
	2	**1.00**	0.41	0.88	0.52	0.65
	3	**1.00**	0.41	0.88	0.52	0.65
	4	**1.00**	0.41	0.88	0.52	0.65
Roth-Erev	1	1.00	0.28	0.92	0.41	0.56
	2	**1.00**	0.28	0.92	0.41	0.56
	3	**1.00**	0.28	0.92	0.41	0.56
	4	**1.00**	0.28	0.92	0.41	0.56
UCB1	1	1.00	0.23	0.76	0.53	0.72
	2	**1.00**	0.23	0.76	0.53	0.72
	3	**1.00**	0.23	0.76	0.53	0.72
	4	**1.00**	0.23	0.76	0.53	0.72
EXP3	1	1.00	0.32	0.63	0.45	0.42
	2	**1.00**	0.32	0.63	0.45	0.42
	3	**1.00**	0.32	0.63	0.45	0.42
	4	**1.00**	0.32	0.63	0.45	0.42

Table 4 shows that the proposed model is able to learn and identify the best scenario for each of the four considered contexts, namely scenario 3 for contexts 1 and 3, and scenario 1 in contexts 2 and 4. On the other hand, all the other state of the art algorithms are able to effectively learn the best global scenario (scenario 1), but, by not including a contextual dimension, they are not able to identify the best scenario for the specific contexts. In summary, the current algorithms are able to learn the best overall approaches, but lack the adaptation capabilities to be able to identify different performances under different contexts.

Table 5 shows that the proposed method is the algorithm that achieves the lowest prediction errors in all four contexts, as result from this method's context aware learning capability. However, some other methods reach very close results in the contexts in which the prediction is from Scenario 1 (identified by all methods as the best one, as seen from Table 4), namely in contexts 2 and 4. Nevertheless, the results from the proposed method are still better in these contexts because it is able to converge faster to the best scenario, by considering the different contexts as independent, while the other methods need for exploration (and more trial and error) to reach the best overall scenario.

Table 5: Comparison of average prediction errors of the different algorithms in each context

Context	Algorithm	MAE	MAPE (%)	STD
1	Proposed Model	7.45	9.89	8.98
	Std. Q-Learning	11.24	16.28	14.04
	Roth-Erev	10.49	14.87	13.41
	UCB1	15.36	21.04	18.93
	EXP3	18.56	24.90	21.39
2	Proposed Model	4.28	6.46	5.89
	Std.Q-Learning	4.88	7.23	6.68
	Roth-Erev	4.46	6.83	6.03
	UCB1	5.89	9.31	8.72
	EXP3	6.53	10.85	9.38
3	Proposed Model	5.37	8.21	6.98
	Std.Q-Learning	9.16	13.28	9.37
	Roth-Erev	8.43	12.73	9.14
	UCB1	12.54	18.02	12. 71
	EXP3	15.11	22.02	17.37
4	Proposed Model	6.22	9.35	8.31
	Std.Q-Learning	6.81	9.97	9.02
	Roth-Erev	7.47	10.28	11.07
	UCB1	6.74	9.63	8.86
	EXP3	7.26	10.15	10.62

The Kruscal-Wallis test is a nonparametric test used to compare three or more independent samples. It indicates if there is a difference between at least two of them. This is used to test the null hypothesis that all populations have equal distribution functions against the alternative hypothesis that at least two of the populations have different distribution functions. In this way it is assumed that equality of averages when equality of equal distributions exists [21]. By the Kruscal-Wallis test it is possible to obtain the value of p = 0 that indicates the rejection of the null hypothesis that all data samples have the same distribution at 1% significance level. The comparison between the pairs of groups is made to verify which of the samples differ from each other.

The Bonferroni procedure is performed to make the comparison in pairs. Fig. 3 represents the 95% confidence interval for all sample groups (5 methods, in which group 1 is the proposed method), in the total of all executions using the three data sets. In this way, it is possible to see which groups differ in the value of the average, using the Bonferroni procedure.

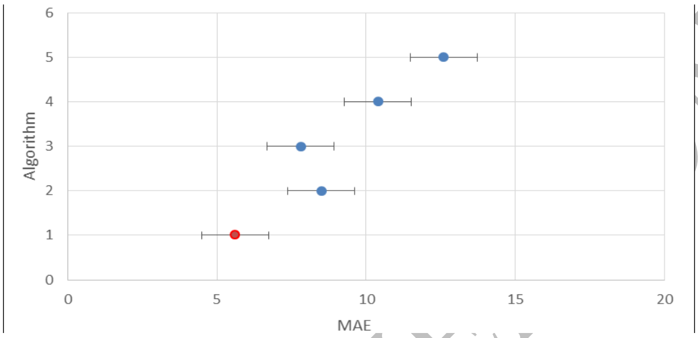


Figure 3: Bonferroni confidence interval by 95%

Fig. 3 shows that all methods have significantly different mean values. Since p = 1 in all these group tests, the null hypothesis where the groups are considered to have similar means with an error of 5% is accepted. Considering this analysis, it is concluded that the applied benchmark methods achieve significantly different results, thus supporting the relevance of the proposed approach.

4 Conclusions

Electricity markets are complex and dynamic environments, involving many entities and a constantly changing negotiation setting. Players acting in this domain need strong decision support solutions in order to be able to take as much benefit from market negotiations as possible.

The model presented in this paper aims at providing decision support to market players by helping them understanding which the best negotiation opponents to negotiate with are. In this way, a learning model for the pre-negotiation stage of bilateral contracts negotiations is presented. This model improved the standard Q-Learning algorithm by including a contextual dimension, thus providing a contextual aware learning model. A simulated annealing process is also included in order to enable accelerating the convergence process when needed, especially when the number of observations is low. Results show that the proposed model is able to undertake context-aware learning, surpassing the results of several benchmark learning algorithms when it comes to contextual learning.

References

1. Lago, J., De Ridder, F., Vrancx, P., & De Schutter, B. (2018). Forecasting day-ahead electricity prices in Europe: The importance of considering market integration. Applied Energy, 211, 890–903
2. Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. Ren & Sust Energy Reviews, 81, 1548–1568
3. Klessmann C, Held A, Rathmann M, Ragwitz M. Status and perspectives of renewable energy policy and deployment in the European Uniondwhat is needed to reach the 2020 targets? Energy Policy December 2011;39(12): 7637e57
4. MISO Energy, homepage, http://www.misoenergy.org (accessed on August 2018
5. NordPool, homepage, http://www.nordpoolspot.com (accessed on August 2018)
6. Soares, J., Pinto, T., Lezama, F., Morais, H. "Survey on complex optimization and simulation for the new power systems paradigm", Complexity, vol. 2018, pp. 32, 2018
7. Pinto, Z. Vale, T. M. Sousa, I. Praça. Negotiation context analysis in electricity markets. Energy 85 (2015) , 78-93
8. Pinto T., Vale Z., Sousa T., Praça I., Santos G. and Morais H. Adaptive Learning in Agents Behaviour: A Framework for Electricity Markets Simulation. Integrated Computer-Aided Engineering, IOS Press, vol. 21, no. 4, pp. 399-415, September 2014
9. Gerber, M., & Bornn, L. (2018). Convergence results for a class of time-varying simulated annealing algorithms. Stochastic Processes and Their Applications, 128(4), 1073–1094
10. Sutton, Barto. 1998. “Reinforcement learning: An introduction”. Reinforcement Learning
11. Rahimi-Kian, A.; Sadeghi, B.; Thomas, R.J. Q learning based supplier-agents for electricity markets. IEEE Power Engineering Society General Meeting, 2005, 1, 420-427
12. Watkins, P. Dayan. 1992. “Q-learning. Machine Learning”. Machine Learning
13. Haznedar, B., & Kalinli, A. (2018). Training ANFIS structure using simulated annealing algorithm for dynamic systems identification. Neurocomputing, 302, 66–74
14. Huang and Y.-H. Hsieh, “Very fast simulated annealing for pattern detection and seismic applications,” Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International. pp. 499–502, 2011
15. Chen, C. Xudiera, and J. Montgomery, “Simulated annealing with thresheld convergence,” Evolutionary Computation (CEC), 2012 IEEE Congress on. pp. 1–7, 2012
16. OMIE. ejecucioncbfom. http://www.omie.es/files/flash/ResultadosMercado.html/, 2018. [Online; accessed March-2019].
17. Pinto, Z. Vale, I. Praça, E.J. Solteiro, F. Lopes. 2015. “Decision Support for Energy Contracts Negotiation with Game Theory and Adaptive Learning”. Energies
18. Erev, I., Roth, A.E.: Multi-agent learning and the descriptive value of simple models. Artif. Intell. 171, 423–428 (2007).
19. Burtini, G., Loeppky, J., Lawrence, R.: A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit. arXiv1510.00757 [cs, stat]. (2015)
20. Bouneffouf, D., Féraud, R.: Multi-armed bandit problem with known trend. Neurocomputing. 205, 16–21 (2016)
21. Theodorsson-Norheim, “Kruskal-Wallis test: BASIC computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples,” Comput. Methods Programs Biomed., 23, 1, 57–62, 1986